JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2025, VOL. 00, NO. 0, 1–6
https://doi.org/10.1080/26939169.2025.2455195

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Identifying Biases and the Relevant Statistical Population: The Case of the Loch Ness Monster

Charles G. M. Paxton[a], Adrian J. Shine[b], and Valentin M. Popov[a]

[a]School of Mathematics and Statistics, Buchanan Gardens, University of St Andrews, St Andrews, Fife, UK; [b]Loch Ness Project, Drumnadrochit, Invernessshire, UK

ABSTRACT

We used a database of Loch Ness Monster reports to show university level students how important it is to think about bias and the identification of the correct statistical population. The Loch Ness Monster reports database illustrates the importance of considering independence, inaccuracy and imprecision when considering data and how statisticians might handle anecdotes as data. Whilst the data is inappropriate for directly making inferences about Loch Ness Monsters, it may be appropriate for making inferences about the population of Loch Ness Monster reports.

ARTICLE HISTORY
Received May 2023
Accepted January 2025

KEYWORDS
Anomalies; Cryptozoology; Monsters

1. Introduction

An important, but often overlooked step in any statistical analysis is the recognition of the appropriate statistical population under consideration. Based on our experience, students (and many scientists!) can be quite sloppy in identification of the correct statistical population. One way to emphasize the importance of identification of the appropriate statistical population is to consider the usage of anecdotes. Anecdotes, here taken to be "the narrative of a detached incident, or of a single event, told as being in itself interesting or striking" (Anonymous 1989), are generally not regarded as amenable to scientific and, by extension, statistical scrutiny. The reasons for this to the statistician are obvious; anecdotes may be inaccurate, subjective or simply false (see Paxton 2022 for an in depth discussion). However, these problems can be a function of the statistical population (i.e., the population we want to make inference about) under consideration relative to the sample. A sample of anecdotes may be an inappropriate sample to make inferences about a particular statistical population but can be an unbiased sample relative to another population (Paxton 2022). For example, since 1930 there have been eyewitness reports of a large unknown animal from Loch Ness, Scotland, (see Williams 2015 for a history). Consider the following account by Greta Finlay in 1952,

> I was so taken up with strange appearance of the head and neck that I did not examine the rest of the animal at all closely. There were two or three humps and the total length visible would be about 15 feet.
>
> The neck was held erect, and where it met the water it enlarged to join a bulky body. The head and neck together were 2-2$\frac{1}{2}$ feet in length, the head alone being about 6 inches long and of about the same width as the neck. What astonished me, apart from the hideous appearance of the head, was that there were two six-inch long projections from it, each with a blob at the end. The skin looked black and shiny and reminded me of a snail more than anything. (Whyte 1961, p. 69)

A written interpreted Loch Ness monster report, like sea monster accounts (Paxton 2009), is the last component of a distinct chain of events. The witness saw (or claimed to see) a monster-like object, he/she decided it was a monster rather than something more prosaic, remembered it and then reported it to someone. The report was then possibly collated and made available to the analyst, who in some way will have engaged in some interpretation/coding prior to analysis. Thus, any statistical analysis of monster reports tells us both about the monster reporting process and perhaps only secondarily the objects actually seen by the witnesses (Paxton 2009, 2010, 2011). The statistical population under consideration here is not Loch Ness Monsters but Loch Ness Monster *recorded and collected reports*. We do not know if there is a population of Loch Ness Monsters but we do know that there is a population of Loch Ness Monster *reports*. In this article we describe how we use the Loch Ness Monster (hereafter LNM) report data for teaching. We use collected reports from Loch Ness as a means to get university level students to think about what questions can be asked of data and what "data" really means. We demonstrate that when the population and the sampling unit are correctly identified, important insights can be gained from anecdotes. In the process, we reveal some interesting features of the LNM report dataset.

CONTACT Charles G. M. Paxton cgp2@st-andrews.ac.uk School of Mathematics and Statistics, Buchanan Gardens, University of St Andrews, St Andrews, Fife, KY16 9LZ, UK.

© 2025 The Author(s). Published with license by Taylor and Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Table 1. Fields used in the Loch Ness Monster data and analysis, total n = 1433.

Name of field	Description	Number of not reporteds	Notes
Report.ID	Report identification number	0	
Event.ID	Event identification number	0	
Anonymous.at.point.of.interview	Whether or not the report was anonymous	0	Levels: Yes/No
Location	Whether witness was on a boat on shore etc.	0	Levels: Shore\Boat\Aerial\Webcam \Not Known
Year	Year of encounter	0	Year as described by witness
Year.for.analysis	Tidied numeric year of encounter	0	Midpoint of plausible range used for imprecisely dated years etc. Range: 1862–2015
EncounterMonthNumeric	Month of encounter	217	Range: 1–12
TimeSI	Time of day as decimal hours that is, 17.5 is 5.30 p.m.	870	Range: 0–23.99
Report.Year.for.analysis	Year of report (can be different to year of encounter)	218	Range: 1896–2016
BSSTranslated	Beaufort wind force scale	987	Range: 0 (flat calm)–10 (storm)
AllDistNADpriority	Estimated reported nearest approach distance of witness to monster (m)	896	Range: 0 – 8449
Duration	Reported (presumably estimated) duration of encounter (minutes)	937	Range: 0.0083–65 mins
SeenLengthSI	Estimated reported seen length of monster (as opposed to estimated total length) (m)	1316	There are other measures of length Range: 0.15–27.43 m

We document several levels of analysis that could be undertaken by first and second year undergraduates or indeed taught to postgraduate students new to statistics. We offer our data and code for others to use.

2. The Data

We created a database of LNM accounts up to 2016 inclusive, made up of a compilation of reports mostly previously published, but including some unpublished accounts from the Mountain expedition to Loch Ness in 1934, reports collected by the Loch Ness Phenomena Investigation Bureau (LNPIB) 1962–1972 and reports collected by the Loch Ness Center. The reports come from personally collected interviews, newspapers, magazines, books and more recently websites and future work will explore the effects of these differing sources. The data and analysis code are available from github (*https://github.com/Tullimonstrum/NessiestatsforJSDSE.git*) and OSF (OSF | Loch Ness Monster Data, *https://osf.io/c3n5f/*).

The database obtained is of a considerable size, in excess of 1800 reports, but the reports are not independent (see the Monstrous Tutorial Section 3). We emphasize that, similar to any other data collection process undertaken to tackle real-world challenges from industry and science, much care was taken in creating this database. In particular, we had clear rules of what reports to include. We considered only reports of phenomena unrecognized by their observers at the time of observation and interpreted as a monster(s). Reports were not considered if they were:

- post-hoc identifications of something anomalous in a photograph where the witnesses did not actually intend to photograph the LNM
- Known hoaxes (however, no attempt was made to distinguish between true or unexposed hoax reports).
- Terrestrial reports of the LNM, for example the famous Spicer sighting of 1933 (Anonymous) and anomalous sonar results
- Reports just consisting of strange noises or smells or where a wake or disturbance was seen but no solid object broke surface or was seen underwater.

The data processing was also done with utmost care. For example, in the spreadsheet where we compiled the data, we entered only the information given by the witness rather than inferred by the reader. The exception was Beaufort wind force scale (an ordinal meteorological scale of wind strength based on observations, where 0 is flat calm, 3 would indicate "white horses" on the waves and 12 is hurricane force winds, with this strength not reported in our observations) which was interpreted from witness' descriptions. The majority of sources were codified for all inputs, at least twice over gaps of sometimes years to ensure the accuracy of the data coding. To use the example given in the previous section, the estimated reported total seen length would be $15 \times 0.3048 = 4.57$ m. The other variables would also be recorded (e.g., reported head and neck etc.) but they are not considered in the analysis here. The full list of variables used in this article is given in Table 1. To the best of our knowledge this is the largest compilation of written reports of the LNM to date. We assume, reasonably, that the thus obtained sample is an unbiased sample of the total population of LNM reports potentially available to us. However, the reports are almost certainly biased with regard to the experienced phenomena, as presumably more vivid and exciting reports are more likely to be preserved. All these points are subject to debate as detailed in the tutorial section below.

3. A Monstrous Tutorial

Our initial approach is to get students, typically those embarking on their statistical learning at undergraduate or master's level, to consider whether they regard anecdotes as data. We have used this material in various introductory classes since 2020, for example as a tutorial in an introductory statistics course (September 2023) for students taking a one-year masters degree (U.K. system). In this case, only 9 out of 29 students initially considered anecdotes could be data. In an undergraduate class (March 19, 2024), 22 out of 71 students of mixed scientific background initially thought anecdotes could be data.

We then got students to debate what would be an appropriate statistical population to consider in discussing the LNM (see

the worksheet in Supplementary Information). The students gradually realize that whatever inference can be made, it cannot be about LNMs as these cannot be sampled. Our sampled population is clearly LNM reports but even that is not the whole story. The population is actually LNM reports after the chain of events described in the Introduction, which involve the witness, the data collector and the analyst. Therefore, the statistical population is actually potential LNM reports accessible to the analyst. We also got the students to consider, through discussion, the biases that could occur at every stage of the report collation process and what the implications would be for making inferences about LNM reports and LNMs.

The reports are almost certainly biased with regard to the experienced phenomena as presumably more vivid and exciting reports are more likely to be preserved. It seems likely our sample was an unbiased sample of the total population of collected LNM reports potentially available to us with regard to the hypotheses we considered, especially as we know of no large untapped source of written reports so the reports considered here must represent a large proportion of the total population of available LNM reports to 2016. However, this point is worthy of debate in class.

Whilst overall, the analyzed report sample is likely to be biased relative to the unrecorded and unpreserved population of reports, individual variables might be biased too. Longer lasting encounters might be preserved more, reports of larger monsters might be preserved relative to reports of little monsters etc.

At the end of the discussion all the students in the above masters group thought the LNM anecdotes could be used as data albeit with caveats. The above undergraduate group were far less convinced with apparently none changing their mind that anecdotes could be data after the discussion (or at least publicly admitting to it)!

The students also considered what the appropriate sampling unit is. They might initially think it is LNMs, but a little more reflection made them realize it must be LNM reports ideally sampled in such a way as to make them independent. Thus, we also need to consider issues of independence. Reports from a LNM encounter may have several witnesses or the witnesses may describe their experience multiple times (see Paxton and Shine (2016) for a statistical analysis). Witnesses will often have discussed their experience amongst themselves, so multiple reports from the same event cannot be considered independent. Considering each report as independent datums would be an example of pseudoreplication. We suggested to the students that the choice of report from each encounter should be carefully thought over. A case could be made for random selection or selection of the report with the most information. The data we supply is filtered such that there is one report per encounter with reports chosen on the basis of the most information. This leaves 1433 nominally independent reports.

The students also considered what the effects of the reporting being first or secondhand might be with the students inferring that secondhand reports might be less accurate (relative to what was originally perceived) compared to first-hand reports. This was tested later on.

Depending on time available, students could sometimes analyze the data on their own (although we provide code here for instructors). The data has also been used in other contexts (computer practicals, exam questions etc.). Either the instructor went through the Monstrous Tutorial sheet in front of the class, or the students worked through it themselves (without code).

The students also reflected on the historical context of the reports. Some spurious claims have been made in the past regarding the sheer amount of LNM reports. Mackal (1976) claimed that there were in excess of 10,000 sightings but gave no evidence to support this claim and only details 272 first and secondhand encounters in his book of which 258 were analyzed. Similarly, Burton (1961) suggested 3000, again without evidence. Why Burton and Mackal suggested such high numbers remains unclear.

4. Exploratory Data Analysis

After discussing the background of the data, we now embark on some exploratory data analysis (EDA, see R code in Supplementary Information and the worksheet). In our classes, the students generate their own code from the questions in the worksheet. EDA can reveal some interesting features and biases in this dataset. There are 1433 encounters from 1850 onwards with 641 first-hand reports. If the most detailed account is taken from each encounter some interesting features are observed. LNMs are normally reported in summer, at midmorning and afternoon and at low Beaufort wind force scale (Figure 1, see code).

Students can consider why more Loch Ness Monster reports occur in the summer than in the winter. Is this a function of the biology of Loch Ness Monsters or the availability of witnesses? Is the dip in reports at human lunchtime a reflection of monster activity or human activity? This pattern is also seen for Champ, the monster of Lake Champlain (Kojo 1992) in North America.

There are many suggested explanations for the Loch Ness Monster (Paxton and Shine 2016). One suggested explanation for some Loch Ness Monster reports is that undulating hump type monsters are caused by wake effects (Gould 1934; Baker and Westwood 1962) where the source of the wake cannot be seen by shore observers close to the water level (Shine, pers. obs.). Such wakes can really only be seen at lower Beaufort wind force scale.

The *median* self-reported duration of encounters was surprisingly long: 4.5 min (range: 0.5 sec to 1 hr 5 min) suggesting that LNMs are not, contrary perhaps to popular belief, generated by misperception or misinterpretation caused by a brief glimpse of an object. This surprisingly long median self-reported duration of the encounters suggests that even given underestimation of time, typically LNM reports are not generated by glimpses and if there is misperception/misinterpretation, it must occur over several minutes. Interestingly the median duration of encounters from our current dataset is four times longer than reported by Willums (1976) with far fewer data. Why this is so is unclear, especially as in our data the duration of encounters in reports has actually decreased over time (see Section 5).

5. Basic Inference

The dataset can be used for basic inference as well, again allowing consideration of potential biases (see the analysis

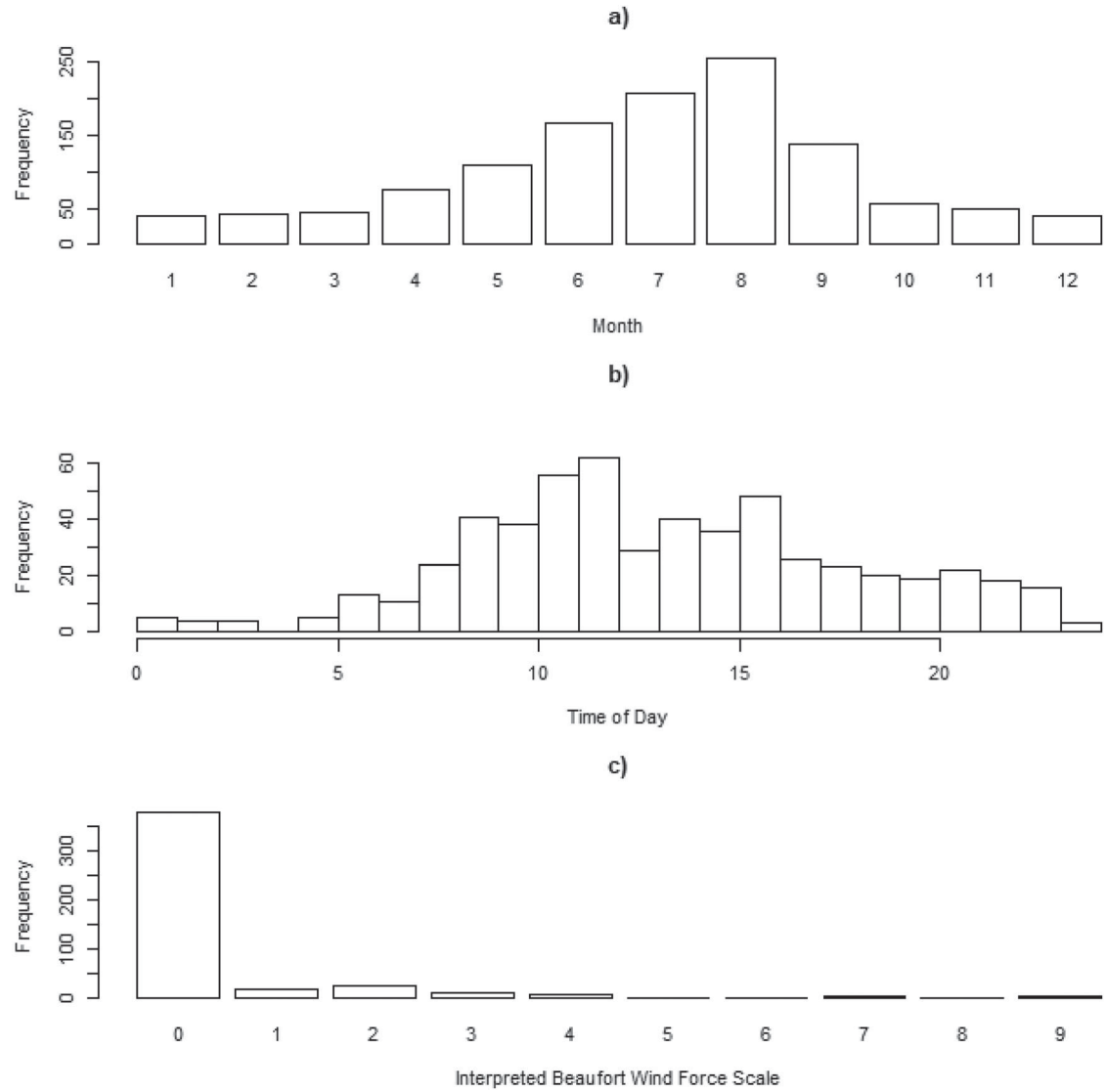

Figure 1. Reported encounters of the Loch Ness Monster by, where known, (a) month of encounter, (b) time of day and (c) interpreted Beaufort wind force scale (low score calm, high score rough waters).

section in the R code file). Students can be encouraged to think about a priori why variables like Anonymity, Location and Handedness might be associated. We can test the hypothesis using χ^2 that Anonymity (i.e., whether or not the witnesses were anonymous) and Location were associated ($\chi^2 = 90.4$, df $= 3$, $p < 0.001$, total n $= 1433$). This was driven by the information content; unknown location reports were often anonymous. Likewise Location and Handedness (whether the reports are first or secondhand) were associated, as information is typically lacking in secondhand reports ($\chi^2 = 121.9$, df $= 3$, $p < 0001$, total n $= 1433$). Note that Location has a "Not Known" level here. Similarly, Anonymity and Handedness were also associated ($\chi^2 = 93.4$, df $= 1$, $p < 0.001$, total n $= 1433$). Again, this association was driven by vague reports that are secondhand with anonymous witnesses. In the above χ^2 analyses there were expected cell sizes of less than 5, but in fact these cells contributed negligibly to the final statistics so no cell amalgamation was undertaken. Loglinear models can also be fitted here to explore higher order interactions.

The reported nearest approach distance of the witness to the monster is related to features of the reports that would lead them to be more unreliable. This can be considered by using regression that initially considered report year, handedness, location and anonymity as potential predictors. The analysis would initially require knowledge of Generalized Linear Models (GLM) because of nonnormal residuals (see analysis of reported nearest approach distance section in code). However, after model reduction and a review of the diagnostics, a nonparametric Kruskal-Wallis test shows there is a difference in reported distance between first- and secondhand reports ($\chi^2 = 5.61$, df $= 1$, $p = 0.018$, $n_{1st} = 321$, $n_{2nd} = 217$). Secondhand reports have a closer reported nearest approach distance than firsthand reports (Figure 2). This could imply some sort of exaggeration in transmission, an interesting example of bias that might creep into reports of other anomalies.

Estimated seen length can also be explored with regression. After model selection, there was evidence for a trend in time (Figure 3). Here we used a GLM with Gamma error (inverse link): ($\chi^2 = 8.082$, df $= 1$, $p = 0.004$, n $= 107$, see analysis

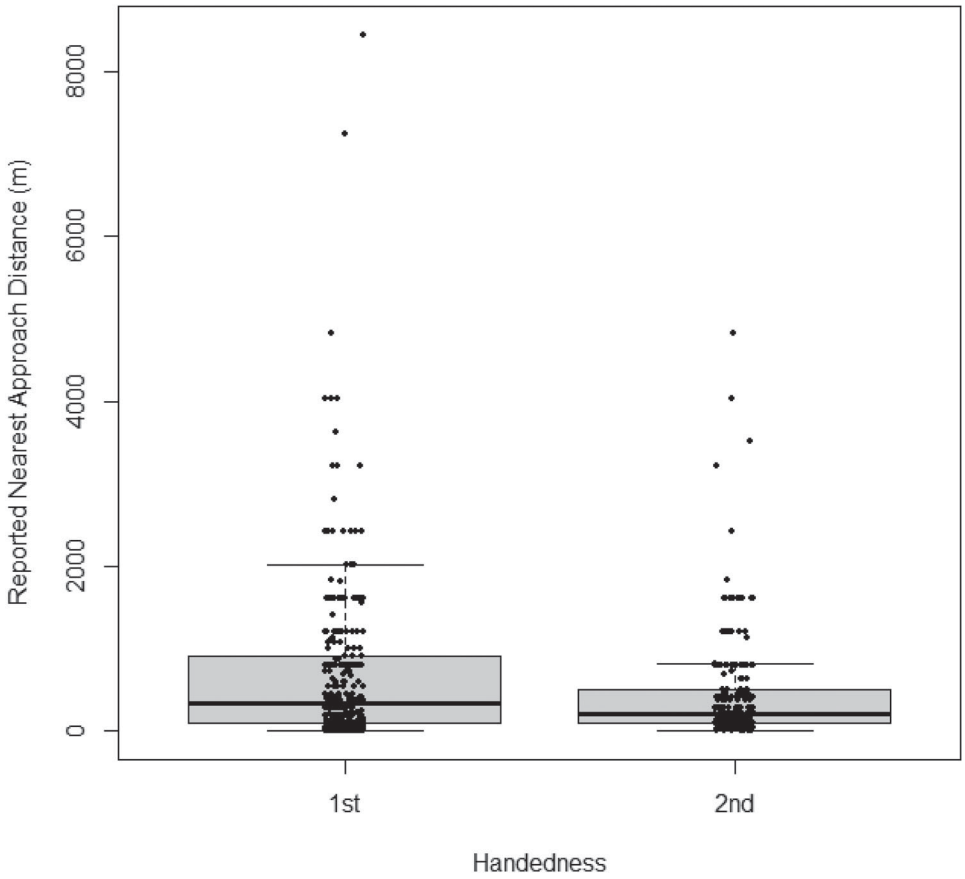

Figure 2. Loch Ness Monster encounter reported nearest approach distances by handedness of report. Horizontal lines represent the nearest lower point to 1.5 times the interquartile range, 25th percentile, median, 75th percentile and the nearest upper point to 1.5 times the interquartile range. Points are jittered to show distribution.

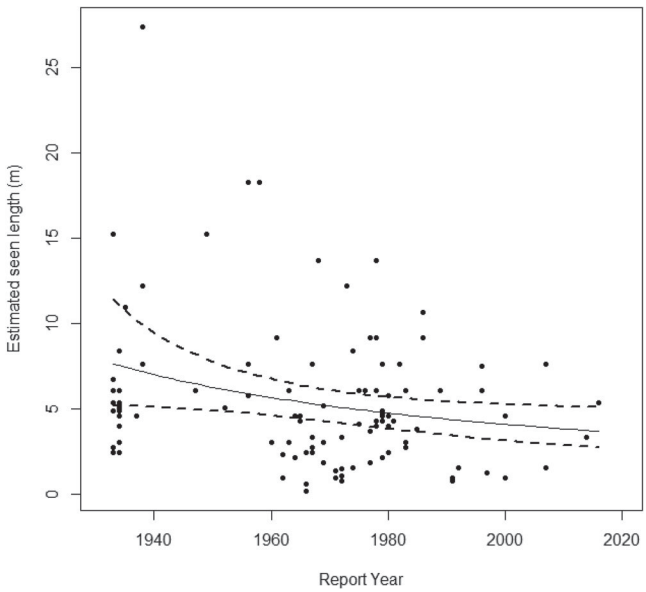

Figure 3. Loch Ness Monster estimated seen length by year of report. The solid line is the best fit line from the Generalized Linear Model with the dashed lines representing 99% confidence intervals.

of estimated seen length in code). The confidence intervals in Figure 3 were constructed using a nonparametric bootstrap, for although the diagnostics generally were fine, there was a hint of a heteroscedasticity issue (see Supplementary Information). The relationship was influenced, however, by an early estimate of great length. Again students may like to speculate why such a trend might exist.

There remain some mysteries in the data with regard to duration. The long reported duration of the encounters has already been commented on. However, the trend in time is quite substantial with an almost 10 min drop in the reported mean duration since the 1930s (GLM with Gamma error (inverse link): $\chi^2 = 26.08$, df $= 1$, $p < 0.001$, n $= 447$, Figure 4, see analysis of duration section in code). Reasons for this trend are unclear and students may wish to suggest how such a trend could arise. Confidence intervals in Figure 4 were obtained by a nonparametric bootstrap because of non-ideal diagnostics (see Supplementary Information).

6. Discussion

The Loch Ness Monster report database is a really useful resource to illustrate biases and the intricacies in identifying the statistical population under consideration. Many of the biases might be plausibly related to distortions caused by the chain of actions required in analyzing the reports. Some results might be reasonably expected: firsthand reports seem less exaggerated than secondhand reports. But some results are genuinely surprising. The self-reported long duration of the encounters is unexpected, as LNMs might be thought to be generated by misidentified brief glimpses of other things. Likewise, it is not obvious why there is a trend in shortening durations through time.

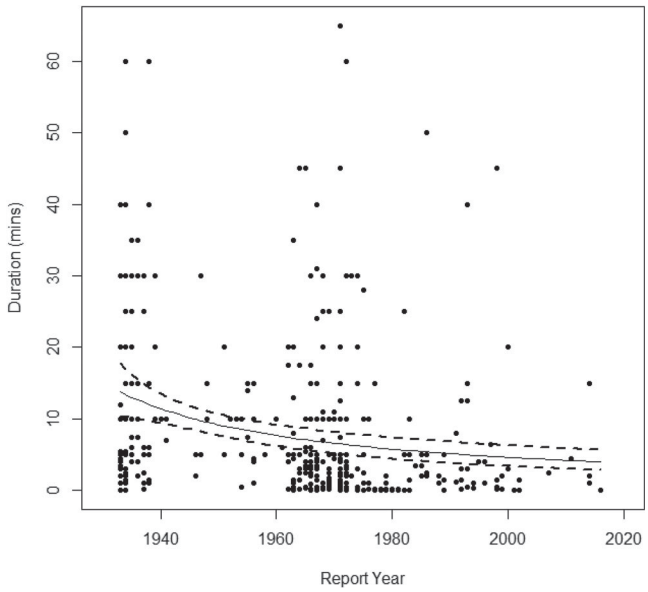

Figure 4. Loch Ness Monster encounter durations by year of report. The solid line is the best fit line with the dashed lines representing 99% confidence intervals obtained from a nonparametric bootstrap.

It is interesting to consider the limits of inference, here. If the sample of reports here represents a large proportion of the "population" of reports is statistics even necessary?

Scientific skeptical texts (e.g., Shermer 2003) often suggest that scientific inferences cannot be made from anecdotes because anecdotes are vague, unsystematic and unrepresentative as indeed they are. Yet many low frequency, albeit important, natural phenomena such as meteorites (Burke 1986), transient atmospheric luminous events (Franz, Nemzek, and Winckler 1990; Boeck et al. 1998) and rogue waves (Draper 1964) have been initially reported only as anecdotes. Therefore, it is worth getting students to consider what inferences can be made from such data as it brings up questions of bias, appropriate statistical populations and appropriate sampling units which are fundamental to any statistical analysis.

Supplementary Materials

Dataset and code.

Acknowledgments

Our thanks to Roland Watson, Gary Campbell and Tony Harmsworth for supplying both some reports and report collection details, Louise Burt for allowing us to sample her class and Jeremy Greenwood, Roland Watson, Henry Bauer and Heidi Burdett for comments on the article. Finally our thanks to a large number of reviewers for comments on the article.

Authors Contributions

Conceptualization: CGMP; Data: CGMP and AJS; Hypotheses generation: AJS, CGMP; Methodology: CGMP, AJS; Visualization: CGMP; Statistics: CGMP and VMP; Writing—original draft: CGMP; Writing—review and editing: CGMP, VMP, AJS.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Blink Films UK provided funding for some of the analyses.

Data Availability Statement

The authors confirm that the data and code supporting the findings of this study are available from github *https://github.com/Tullimonstrum/NessiestatsforJSDSE* and OSF *https://osf.io/c3n5f/*.

ORCID

Charles G. M. Paxton https://orcid.org/0000-0002-9350-3197
Valentin M. Popov https://orcid.org/0000-0003-1079-7992

References

Anonymous. (1989), *Oxford English Dictionary*, Oxford: Oxford University Press.
Baker, P., and Westwood, M. (1962), "Sounding Out the Monster," Observer.
Boeck, W. L., Vaughan, O. H., Jr., Blakeslee, R. J., Vonnegut, B., and Brook, M. (1998), "The Role of the Space Shuttle Videotapes in the Discovery of Sprites, Jets and Elves," *Journal of Atmospheric and Solar-Terrestrial Physics*, 60, 669–677.
Burke, J. (1986), *Cosmic Debris: Meteorites in History*, Berkeley, CA: University of California Press.
Burton, M. (1961), *The Elusive Monster*, London: Hart-Davis.
Draper, L. (1964), "'Freak' Ocean Waves," *Oceanus,* 10, 13–15.
Franz, R. C., Nemzek, R. J., and Winckler, J. R. (1990), "Television Image of a Large Upward Electrical Discharge Above a Thunderstorm System," *Science*, 249, 48–51. DOI:10.1126/science.249.4964.48
Gould, R. T. (1934), *The Loch Ness Monster*, London: Geoffrey Bles.
Kojo, K. (1992), "Distributional Patterns of Cryptid Eyewitness Reports from Lake Champlain, Loch Ness, and Okanagan Lake," *Cryptozoology*, 11, 83–89.
Mackal, R. P. (1976), *The Monsters of Loch Ness*, Chicago: The Swallow Press.
Paxton, C. G. M. (2009), "The Plural of 'Anecdote' Can be 'Data': Statistical Analysis of Viewing Distances in Reports of Unidentified Large Marine Animals 1758–2000," *Journal of Zoology*, 279, 381–387.
Paxton, C. G. M. (2010), "The Monster Manual," Fortean Times, 54.
Paxton, C. G. M. (2011), "Putting the "Ology" into Cryptozoology," *Bio-fortean Notes,* 1, 7–20.
Paxton, C. G. M. (2022), "Are Skeptics Wrong About the Value of Anecdotes? *Skeptical Inquirer,* 46, 53–57.
Paxton, C. G. M., and Shine, A. J. (2016), "Consistency in Eyewitness Reports of Anomalies: Examples from Accounts of Aquatic "Monsters"," *Journal of Scientific Exploration,* 30, 16–26.
Shermer, M. (2003), "Show Me the Body," *Scientific American*, 288, 37–37.
Whyte, C. (1961), *More Than a Legend*, London: Hamish Hamilton.
Williams, G. (2015), *A Monstrous Commotion*, London: Orion Books.
Willums, J.-O. (1976), "A Computer Analysis of Loch Ness Data," in *The Monsters of Loch Ness*, ed. R. P. Mackal, pp. 341–348, London: Futura.